Behavioral/Systems/Cognitive

A Biophysically Based Neural Model of Matching Law Behavior: Melioration by Stochastic Synapses

Alireza Soltani and Xiao-Jing Wang
Volen Center for Complex Systems and Department of Physics, Brandeis University, Waltham, Massachusetts 02454

In experiments designed to uncover the neural basis of adaptive decision making in a foraging environment, neuroscientists have reported single-cell activities in the lateral intraparietal cortex (LIP) that are correlated with choice options and their subjective values. To investigate the underlying synaptic mechanism, we considered a spiking neuron model of decision making endowed with synaptic plasticity that follows a reward-dependent stochastic Hebbian learning rule. This general model is tested in a matching task in which rewards on two targets are scheduled randomly with different rates. Our main results are threefold. First, we show that plastic synapses provide a natural way to integrate past rewards and estimate the local (in time) "return" of a choice. Second, our model reproduces the matching behavior (i.e., the proportional allocation of choices matches the relative reinforcement obtained on those choices, which is achieved through melioration in individual trials). Our model also explains the observed "undermatching" phenomenon and points to biophysical constraints (such as finite learning rate and stochastic neuronal firing) that set the limits to matching behavior. Third, although our decision model is an attractor network exhibiting winner-take-all competition, it captures graded neural spiking activities observed in LIP, when the latter were sorted according to the choices and the difference in the returns for the two targets. These results suggest that neurons in LIP are involved in selecting the oculomotor responses, whereas rewards are integrated and stored elsewhere, possibly by plastic synapses and in the form of the return rather than income of choice options.

Key words: **matching behavior; reward-dependent stochastic Hebbian learning; lateral intraparietal cortex; melioration; decision making; dopamine**

Introduction

In natural behavior, what often matters is how we make a series of choices over time, rather than an isolated decision. For example, the success in foraging for nourishment depends on the temporal pattern of food gathering; one's diet is determined by how frequently one selects food alternatives over a period of time. For decades, psychologists have studied individuals' allocation of repeated responses to a set of choices in laboratory experiments using foraging-type tasks. In these tasks, the environment is uncertain and the same choice can lead to different outcomes (no reward, or reward of varying magnitude); thus, decision making is inherently probabilistic. These studies have led to Herrnstein's "matching law," which states that a subject allocates her or his choices in a proportion that matches the relative reinforcement obtained on these choices (Herrnstein, 1961; Williams, 1988; Herrnstein et al., 1997). The matching law has been shown to be valid in a variety of task paradigms, and across species (e.g., pigeons, rats, monkeys, humans) (de Villiers and Herrnstein, 1976; Williams, 1988; Gallistel, 1994; Anderson et al., 2002).

Matching law is about an individual's choice, hence ultimately should be explained in terms of neural processes of decision making in the brain. Recently, neurobiologists have embarked on this quest and have begun to identify single neuronal activities in the primate brain that are correlated with matching behavior. In particular, several studies have used oculomotor tasks, in which typically two visual targets for saccadic eye movements are associated with different probabilities and/or magnitudes of rewards (Platt and Glimcher, 1999; Sugrue et al., 2004a; Lau and Glimcher, 2005b). Platt and Glimcher (1999) found that spike firing of single cells in the lateral intraparietal cortex (LIP) (a cortical area critical to oculomotor behavior) varies with the relative gain that an animal expects from each response, as well as with the probability of obtaining such a reward. Therefore, it was suggested that the LIP activity is modulated by relative profitability (expected gain times reward frequency). Sugrue et al. (2004a) used concurrent variable-interval schedules similar to the original Herrnstein experiment, and found that activities of some LIP neurons were correlated with a representation of value that the authors defined as fractional income. This study along with other studies (Platt and Glimcher, 1999; Dorris and Glimcher, 2004) indicates that LIP neurons reflect the values of possible actions, although these values are likely to be computed elsewhere in the brain [LIP neurons are selective to the spatial location of a visual target, whereas in this experiment the value (baiting probability) is associated with target color rather than its location]. Moreover, phenomenological models, in which the local (response by response) decision is based on time integration of past rewards (Sugrue et al., 2004a; Corrado et al., 2005b) or both past rewards and choices

Received Dec. 3, 2005; revised Jan. 26, 2006; accepted Feb. 17, 2006.

This work was supported by National Institutes of Health Grants MH062349 and MH073246 and by the Swartz Foundation. We are grateful to Stefano Fusi for the early work and helpful discussion on stochastic learning and to L. Sugrue for helpful discussions on the matching experiment. We also thank Z. Ayubi, R. Gallistel, B. Lau, D. Luxat, and P. Miller for helpful comments on this manuscript.

Correspondence should be addressed to Xiao-Jing Wang at the above address. E-mail: xjwang@brandeis.edu.

DOI:10.1523/JNEUROSCI.5159-05.2006

(Lau and Glimcher, 2005b) have been shown to reproduce monkeys' matching behavior. These and other models (Williams, 1988; Gallistel et al., 2001), however, do not address the question of what cellular and circuit mechanisms, at the biophysical level, underlie the matching behavior, which is the subject of the present work.

Our starting point is a biophysically based spiking neuron model of decision making (Wang, 2002) that has been shown to capture psychological behavior and corresponding LIP neural activities in perceptual (visual motion) discrimination tasks (Shadlen and Newsome, 1996, 2001; Roitman and Shadlen, 2002). In that model, two groups of neurons (tuned to different targets) integrate inputs over time, and the choice is selected according to which of the two neural groups wins the competition.

In the present study, we incorporated reward-dependent synaptic plasticity into our neuronal decision-making model. Specifically, we used binary synapses that undergo a stochastic Hebbian learning rule (Amit and Fusi, 1994; Fusi, 2002), with the additional condition that coactivation of presynaptic and postsynaptic neurons leads to potentiation only if the choice is rewarded, and depression otherwise. This was inspired by the suggestion that the presence or absence of dopamine signal modulates the synaptic plasticity at corticostriatal and prefrontal synapses (Reynolds et al., 2001; Reynolds and Wickens, 2002; Jay, 2003; Otani et al., 2003; Huang et al., 2004). Our working hypothesis is that input synapses onto a decision circuit (like LIP) are updated at the end of each trial according to such a reward-dependent Hebbian learning rule. As a result of synaptic modifications, the difference in the input strengths for the two competing neural groups of the decision network varies from trial to trial, which leads to adaptive dynamics of choice behavior.

Our model endowed with plastic synapses is a general one, not designed specifically for a particular behavioral task. In this paper, we report model simulations in which the two competing choices were rewarded stochastically at different rates, like in a matching task (Sugrue et al., 2004a; Lau and Glimcher, 2005b). We found that the model reproduces the neurophysiological as well as behavioral observations from the monkey experiment (Sugrue et al., 2004a). We show that plastic synapses provide a natural mechanism for computing local returns (local time average of reward per choice). Moreover, the model operates in single trials according to the so-called melioration principle (i.e., in each trial, the decision is biased toward the choice with a higher return) (Herrnstein and Vaughan, 1980; Vaughan, 1981; Herrnstein and Prelec, 1991), which ultimately gives rise to the global matching behavior. Some preliminary results have been reported in abstract form (Soltani and Wang, 2004).

Materials and Methods

Decision-making network. The decision-making network model used here is the same as the one in the study by Wang (2002); all of the model details can be found therein (see also Brunel and Wang, 2001). Briefly, the model consists of 2000 integrate-and-fire (1600 excitatory and 400 inhibitory) neurons, which are organized into three populations of excitatory neurons (two selective for competing targets, A and B, whereas the third one is nonselective) and one single population of inhibitory neurons. Recurrent synaptic currents are modeled by realistic kinetics, mediated by AMPA and NMDA receptors for excitation and by GABA_A receptors for inhibition. In addition to recurrent synaptic inputs from other neurons in the network, every neuron receives independent background and afferent excitatory inputs (mediated by AMPA receptors) from 800 presynaptic neurons outside the network. The background presynaptic neurons fire constantly at 3 Hz, and these external spikes are generated with Poisson statistics.

With the presentation of visual targets, neurons in both selective populations receive a combination of two inputs (mediated by AMPA receptors). The first one, through a feedforward sensory pathway, codes the target appearance, whereas the second one codes the target color via an indirect pathway. Specifically, the first input (identical for both targets) is mediated by some afferent neurons (eight presynaptic neurons for each neuron in the selective populations), which after the appearance of the two targets, increase their firing rates from 3 to 55 Hz. Moreover, these neurons exhibit spike-frequency adaptation with a time constant of 120 ms and a steady-state firing rate of 8 Hz. In this way, neurons in the two selective populations display an initial peaked response followed by a decay to a lower steady-state response, similar to the response of LIP neurons after the onset of two targets (Sugrue et al., 2004a). The second input is mediated by some other afferent neurons (four presynaptic neurons for each neuron in the selective populations), which increase their firing rates from 1 to 10 Hz during the target presentation. Because in the experiment of Sugrue et al. (2004a), rewards were associated with the choice about color (red and green), we assume that synapses of the second pathway are endowed with reward-dependent plasticity. Because the second input is presumed to arrive through an indirect pathway passing several synapses, we added a latency of 50 ms between the onset of the first and second inputs (Schmolesky et al., 1998).

In the experiment of Sugrue et al. (2004a), the target color (red and green) encodes the rewarding value of each target, but the location of red and green targets was randomized from trial to trial. So in order for spatially selective LIP neurons to receive the correct information about the rewarding value of the leftward and rightward targets, a remapping from color to location should take place in each trial. The model presented in this paper does not explicitly address the issue of remapping from color to location (because we do not have enough experimental information yet on which to build a realistic model), but in the supplemental material (available at www.jneurosci.org), we describe a schematic circuit that is potentially able to perform such a remapping. Regardless of the details of implementation, what is essential for our model is the assumption that competing neural populations in the decision-making network are selective to target options (A or B) and receive inputs that convey information about the associated rewards via plastic synapses.

For the sake of simplicity, we did not model an additional network that reads out the decision choice. Instead, we assumed that 1.2 s after the target onset, if the difference between the average firing rates of the two (A and B) selective populations exceeded a fixed threshold of 8 Hz (for an interval which lasts >50 ms), then the choice of the network was the population with a higher firing rate. In the rare situation when this criterion was not met until 1.5 s after the trial onset, the threshold for making a decision was lowered to 4 Hz. After the decision is made, a decrease in the input firing rates brings the network to a regime that only one of the selective populations can stay at a high level of activity, so at the time of reward delivery the activity in only one of the selective populations is high.

Reward-dependent plasticity rule. The plastic synapses of the second input pathway are assumed to be binary (Petersen et al., 1998; O'Connor et al., 2005), with two discrete states: a potentiated "Up" state with peak conductance of $g_+ = 5.5$ nS, and a depressed "Down" state with peak conductance of $g_- = 0.5$ nS. At any moment, a fraction c of these synapses are in the Up state, whereas the remaining fraction, $1 - c$, are in the Down state. Plasticity is implemented by activity-dependent modifications of c_A and c_B for the two selective and competing neural populations.

The learning rule we use has three characteristics (Fusi et al., 2005a). First, it is Hebbian, depending on the firing rates of presynaptic (target-coding) and postsynaptic (decision) neurons. Second, an all-or-none reward signal (depending on the outcome of a target selection) can reverse the direction of plasticity (potentiation if reward is harvested, depression otherwise). Third, when the Hebbian condition is met, synaptic modification occurs probabilistically (Amit and Fusi, 1994; Fusi, 2002; Fusi et al., 2005b). In potentiation instances, each synapse in the Down state has a probability q_+ to be switched to the Up state. Similarly, in depression instances, each synapse in the Up state has a probability q_- to be switched to the Down state.

Based on these rules, the fraction of synapses in the Up state, c_i, is updated at the end of each trial as follows:

$$c_i(n+1) = c_i(n) + q_+(r;\,\nu_i)[1 - c_i(n)] \text{ in the case of LTP} \tag{1}$$

$$c_i(n+1) = c_i(n) - q_-(r;\,\nu_i)c_i(n) \text{ in the case of LTD,} \tag{2}$$

where i = A or B, and $q_+(r;\,\nu_i)$ and $q_-(r;\,\nu_i)$ are the potentiation and depression rates, respectively (termed together as learning rates). The second term in Equation 1 describes the change attributable to the transition of synapses in the Down state, because a fraction $1 - c_i$ of synapses are potentiated with probability $q_+(r;\,\nu_i)$. Similarly, the second term in Equation 2 describes the change attributable to the transition of synapses in the Up state, because a fraction c_i of synapses are depressed with probability $q_-(r;\,\nu_i)$. The learning rates depend on the firing rate ν_i, of the postsynaptic decision neurons at the end of each trial, and on the outcome of the decision r, which is either rewarded or unrewarded. The firing rate ν_i is low for the neurons selective to the unchosen target, and it is high for the neurons selective to the chosen target. For most of the results presented in this paper, unless stated otherwise, we assume that the depression and potentiation rates are constant and nonzero if ν_i is high, so synaptic plasticity only happens to the set of synapses projecting to the winning neural population. In this case, the learning rule simplifies to the following:

$$\begin{aligned} c_i(n+1) &= c_i(n) + q_+[1 - c_i(n)] \text{ target } i \text{ is selected and rewarded} \\ c_i(n+1) &= c_i(n) - q_-c_i(n) \text{ target } i \text{ is selected but not rewarded.} \end{aligned} \tag{3}$$

Matching task simulation. An oculomotor matching task paradigm similar to that of Sugrue et al. (2004a) was simulated. In this task, a monkey was trained to choose between two visual targets with different colors (red and green). A selection of each target is rewarded independently and stochastically at a certain rate (with Poisson statistics). Reward in this task was persistent in the sense that, if a reward was assigned to a target, it stayed there until it was harvested. To discourage the monkey from switching between the two targets, a change-over-delay (COD) penalty was imposed, so if the monkey switched from one target to the other, it should choose the new target for the second time to harvest any baited reward on it. The probability of baiting rewards on the two targets (reward schedule) changed between blocks of trials without any warning to the monkey. The baiting probability ratios were randomly chosen from the ratios [1:1, 1:3, 1:6, 1:8], whereas the overall baiting probability was fixed.

In our study, we use a discrete version of the same task so in each trial if a target was not baited with a reward, the computer assigned a reward to that target with some probability, independently of the other target. The overall baiting probability is set to 0.3 rewards per trial to match the reward rate in the experiment of Sugrue et al. (2004a). A sequence of blocks with different baiting probability ratios is called a "session." In most of the simulations, the model encountered a session of the matching task with baiting probability ratios [1:1, 1:3, 3:1, 1:1, 3:1, 1:3, 1:1, 1:6, 6:1, 1:1, 6:1, 1:6, 1:1, 1:8, 8:1, 1:1, 8:1, 1:8, 1:1], which were presented in a sequence of blocks of trials. This reward schedule was usually fixed when the performance of the model was assessed with a range of parameter values. In this way, the model has been tested with the most drastic changes in the reward schedule. The average choice behavior of the model is then computed using all blocks of trials. As observed in the experiment, monkeys obey the COD constraint so most of the time they stay on the new selected target after a switch. In our modeling, we impose the COD by requiring the model to choose the same target after any switch. Similar to other trials, in the trials after switches, plastic synapses undergo changes according to the same learning rule. In one simulation (see Fig. 6), we relax the COD constraint so in that case there is no mandatory movement after a switch and the model freely chooses between the two targets in each trial.

It is important to define the following terms that we use throughout the paper. If from the total number of N trials, N_A of them were choices for target A and N_B were choices for target B, then the probability of choosing A, P_A, is equal to N_A/N. At the same time, if the total M_A and M_B rewards have been harvested on target A and B, then the incomes from target A and B, I_A and I_B, are equal to M_A/N and M_B/N, respectively. Furthermore, returns from target A and B, R_A and R_B, are equal to M_A/N_A and M_B/N_B, respectively.

Results

Behavior of the decision-making network and emergence of graded activity

The behavior of our model results from an interplay between the decision process and synaptic plasticity. In any trial, given the synaptic strengths c_A and c_B, the network integrates inputs and makes a choice. At the end of each trial, depending on the choice and whether it is rewarded, c_A or c_B is updated, which in turn influences the decision process in the subsequent trial. We first quantify the decision process of the network as a function of fixed c_A and c_B values. Then we will consider trial-to-trial modifications of c_A and c_B and the resulting dynamic decision making over time.

In our simulation of the experiment of Sugrue et al. (2004a), it is reasonable to assume that the two visual targets lead to identical firing rates of presynaptic sensory neurons that project to our decision network. Therefore, the only difference in the inputs to the two competing neural populations is attributable to the efficacies of the plastic synapses. The average synaptic conductance of input plastic synapses is a function of multiple factors, including the number of plastic synapses, the presynaptic firing rate, and the peak conductance of the potentiated and depressed states, and can be written as follows:

$$G = N_p f_{st}(cg_+ + (1-c)g_-)\tau_{syn}, \tag{4}$$

where N_p is the number of plastic synapses onto each neuron, f_{st} is the firing rate of the presynaptic neurons, g_+ and g_- are the peak conductance of the synapses in the Up and Down states, respectively, and τ_{syn} is the decay time of AMPA currents. Thus, the difference in the average synaptic conductances of neurons in the two selective populations, can be quantified as a function of the synaptic strengths (fraction of synapses in the potentiated state) c_A and c_B.

$$G_A - G_B = (c_A - c_B)N_p f_{st}(g_+ - g_-)\tau_{syn}. \tag{5}$$

As we show later, the choice behavior of the network is a function of the difference in synaptic strengths $c_A - c_B$ (or equivalently $G_A - G_B$), so the multiplicative factor $N_p f_{st}(g_+ - g_-)$ can change the sensitivity (which we will call σ) of the choice behavior to the difference in synaptic strengths.

The behavior of the network in 10 sample trials, with fixed $c_A = 0.33$ and $c_B = 0.27$, is illustrated in Figure 1. As is evident, at the onset of the targets, the firing rates of the two neural populations initially increase together for a few hundred milliseconds, and then start to diverge so that firing rate of one population (e.g., A) keeps increasing while the firing rate of the other population (e.g., B) decreases gradually.

This "winner-take-all" competition process is attributable to effective mutual inhibition (through the shared interneuron pool) between neural populations A and B. Consequently, at the end of a trial, a categorical choice can be read out according to which of the two neural populations has a higher firing rate. In the model, the high level of activity in the winning population can be self-sustained, even after the removal of the stimulus, because of recurrent reverberation. Hence, in principle, the choice can still be read out after a memory delay period (Wang, 2002).

If the difference between c_A and c_B is not too large, the decision of the network is probabilistic, because neural spike discharges are intrinsically stochastic (note the trial-by-trial variability of population firing rates in Fig. 1). For example, with $c_A = 0.33$ and $c_B = 0.27$, the network chooses A in 78% of trials and B in 22% of trials (sample trials are shown in Fig. 1). A comparison between the left and right panels in Figure 1 reveals that a few hundred milliseconds after the onset of the targets, the firing rate of the winning population is somewhat lower when its synaptic strength is smaller. Moreover, the time it takes for the two neural populations to diverge (hence the "decision time") is longer and more variable.

The trial-averaged population activities are shown in Figure 2*A* for which the average activity in the two selective populations are sorted according to the choice of the network in each trial. It is apparent that a few hundred milliseconds after the onset of the targets, the activity of neurons is significantly higher when the chosen target is the preferred target (red) than when it is the nonpreferred target (blue). Furthermore, there is a graded change of activity levels between the cases in which the synaptic strength for the winning population is larger (thick curves) or smaller (thin curves) than that for the losing population.

These characteristics are robust in our model, as long as the overall synaptic strength ($c_A + c_B$) is reasonably low, which favors winner-take-all competition. As shown in Figure 2, *B* and *C*, with the same value for the difference $c_A - c_B = 0.06$ but a higher value for $c_A + c_B$ (1 and 1.4 instead of 0.6), the winning and losing neural populations still exhibit some differences that are no longer of a categorical character. This is because the external drive is now large for both neural populations and becomes predominant over the winner-take-all recurrent network dynamics, so the activity of the losing population is larger and closer to that of the winning population. In this case, the choice is determined in simulations by the neural population with a higher firing rate at the end of each trial (see Materials and Methods).

Interestingly, for all three cases in Figure 2 with $c_A - c_B = 0.06$, the choice probability turns out to be approximately the same ($P_A = 0.78$, 0.81, and 0.77, for $c_A + c_B = 0.6$, 1, and 1.4, respectively). This holds true for other $c_A - c_B$ values. As shown in Figure 3, the choice probability as a function of $c_A - c_B$ is not sensitive to the overall synaptic strength $c_A + c_B$. We fitted the probability of choosing target A by a sigmoid function of $c_A - c_B$.

$$P_A = \frac{1}{1 + \exp\left(-\frac{c_A - c_B}{\sigma}\right)} \tag{6}$$

Note that the σ value, which determines the randomness in the choice behavior of the network, depends on factors that can change the difference in the overall synaptic currents through the plastic synapses (Eq. 5).

As a result, the value of σ can be adjusted by the number of the plastic synapses, by the firing rates of presynaptic neurons projecting to the decision network, and by the peak conductance of plastic synapses (Eq. 5). For model parameters used in our simulations, we obtained σ equal to 4.84% (if $c_A - c_B$ is expressed as percentage).

It is an important feature of our model that the choice behavior is only a function of the difference ($c_A - c_B$), which we will refer to as "differential input." In Figure 4, the average population activities are sorted according to the choice of the network and the differential input value in each trial when a range of differential input values are used in different trials. The overall synaptic strength ($c_A + c_B$) in these simulations varies from 0.4 to 1.6. Similar to what is shown in Figure 2, a graded activity emerges, which is a direct result of competition in the decision-making network and difference in synaptic strengths of inputs to the two populations.

The activity of neurons, in trials in which they win the competition (the chosen target is their preferred one), is higher when the difference in the synaptic strength in their favor is greater (compare the red curves). In contrast, in trials in which they lose the competition (the chosen target is the nonpreferred one), the activity of neurons is higher when the difference in the synaptic strength in their disfavor is smaller (compare the blue curves).

Therefore, by sorting neural firing rates across trials according to the differential input graded activities emerge in our model. This plot (Fig. 4) is similar to Figure 5 in the study by Sugrue et al.

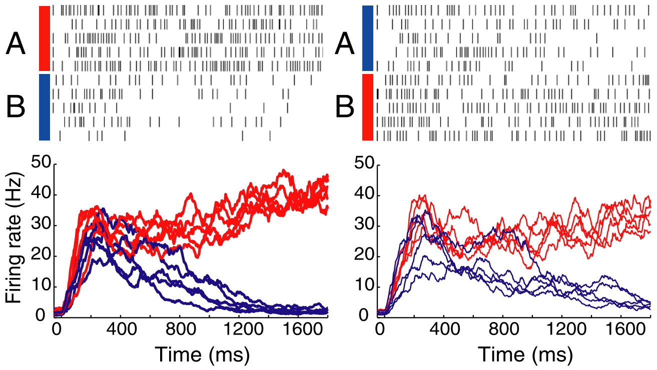

Figure 1. Neuronal activity of two selective populations of decision-making network model, in sample trials. The population firing rate of neurons is shown separately for trials in which the choice of the network is the preferred (red) or nonpreferred (blue) target of the neurons. Raster plots show spike trains for two selected neurons in populations A and B. The left panels show activity in trials in which target A is the choice of the network, and the right panels show activity in trials in which target B is the choice of the network. Activity is aligned at the onset of the visual targets. A few hundred milliseconds after the input onset, the average firing rates in the two populations start to diverge. Spiking activity is higher when the chosen target is preferred for the neuron (compare red with blue traces) and when its input is larger (compare red traces in the left and right panels). Moreover, firing activity is higher when the chosen target is nonpreferred for the neuron that receives a larger input (compare blue traces in the left and right panels). In these simulations, the synaptic strengths are $c_A = 0.33$ and $c_B = 0.27$.

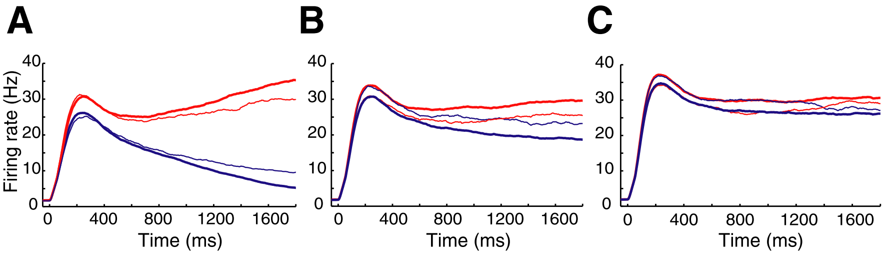

Figure 2. Average activity of the two selective populations sorted according to the choice of the network in each trial. Average activity is shown for three different sets of synaptic strengths: $c_A = 0.33$, $c_B = 0.27$ (***A***); $c_A = 0.53$, $c_B = 0.47$ (***B***); and $c_A = 0.73$, $c_B = 0.67$ (***C***). Note that, in these three cases, the differential input is the same ($c_A - c_B = 0.06$), but the overall inputs are different ($c_A + c_B = 0.6$, 1.0, and 1.4, respectively). The average activity of neurons in the two selective populations are sorted based on the choice of the network in each trial and then averaged (over 400 trials for each set of synaptic strengths). Red (blue), The choice of the network is the preferred (nonpreferred) target for the neural population. Thick (thin), The neural population selective for the chosen target receives a larger (smaller) input than its competitor. Regardless of whether the chosen target is preferred (red curves) or nonpreferred (blue curves), the average population activity is higher when the neurons receive a stronger input (compare thick and thin curves).

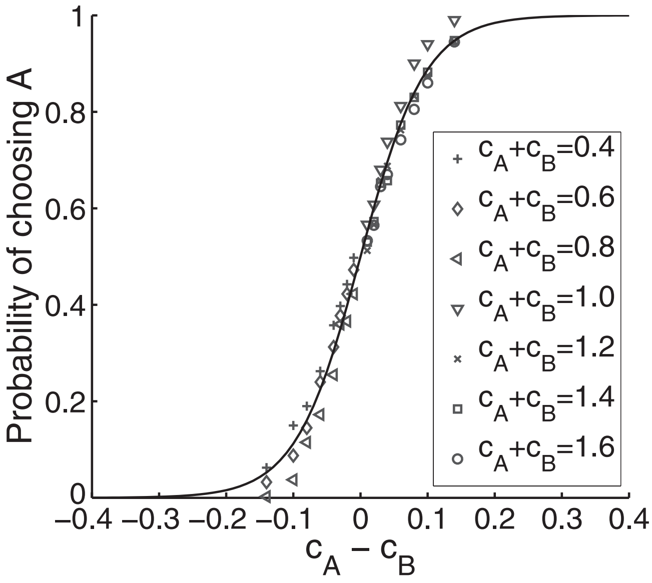

Figure 3. Choice behavior of the decision-making network for different sets of synaptic strengths. The probability of choosing target A is only a function of the difference between the two synaptic strengths, and it can be fitted by a sigmoid function. The solid curve shows the fitting by a sigmoid function ($\sigma = 4.84\%$). The choice probability for each set of synaptic strengths is obtained from 400 simulated trials. For each set of synaptic strengths with equal overall synaptic strengths, differential inputs are set to 0.01, 0.02, 0.03, 0.04, 0.06, 0.08, 0.1, and 0.14.

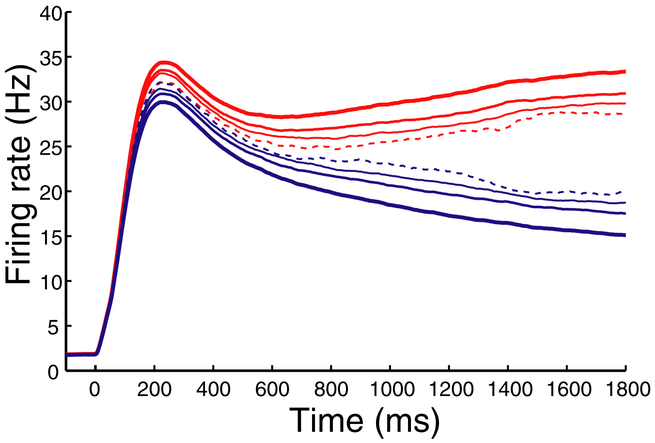

Figure 4. Graded activity of neurons in the two selective populations. The activity of decision neurons shows a graded pattern if single-trial firing rates are sorted and averaged according to the choice of the network and the difference between synaptic strengths. Activity is aligned by the onset of two targets, and it is shown separately for the choice that is the preferred (red) or nonpreferred (blue) target of the neurons. In addition, trials are subdivided into four groups according to the difference between the strength of synapses onto the two competing neural populations [$c_A - c_B = -0.05$ to -0.14 (dashed), 0 to -0.05 (thin), 0–0.05 (normal), 0.05–0.14 (thick)]. The overall synaptic strength, $c_A + c_B$, varies from 0.40 to 1.6. For these simulations, 56 different sets of synaptic strengths are used (the same values used for Fig. 3), and for each set of synaptic strengths the average activity is obtained from 400 simulated trials.

(2004a), in which LIP neural activities are sorted according to the fractional income of the chosen target, which according to the model used in that study is identical to the choice probability. Therefore, our model semiquantitatively reproduces the LIP neural spike activities reported in the matching task (Sugrue et al., 2004a). Because the choice probability P_A (the probability of choosing target A) is a sigmoid function of differential input (Eq. 6), the same kind of graded activities are expected when the neural activities are sorted according to P_A (results are not shown). We shall come back to possible implications of this result to the interpretation of the observed LIP neural activities in Discussion.

The disparate timescales of the neural firing dynamics (milliseconds) and synaptic plasticity (many trials) made it difficult to simulate the full large-scale network model of spiking neurons (which requires a time resolution of 0.1 ms) sequentially across thousands of trials. To avoid this computational hindrance, from now on we use the function $P_A(c_A - c_B)$ (Eq. 6) instead of the full spiking neural network for decision computations. Namely, in each trial, knowing c_A and c_B, we use $P_A(c_A - c_B)$ to flip a biased coin; the outcome of the coin toss determines the choice of the network in that trial (A or B). At the end of each trial, depending on the choice of the model and the presence or absence of reward, plastic synapses undergo stochastic strengthening or weakening according to Equation 3.

Learning rule and the steady state of plastic synapses

In the last section, we quantified the choice behavior of the decision-making network as a function of the differential input, $c_A - c_B$, to the two competing neural populations. Now, we consider the learning process in which c_A and c_B undergo changes depending on the decision of the network and the outcome of that decision (rewarded or unrewarded) at each trial. We first asked the question: for a given and fixed P_A (and $P_B = 1 - P_A$), how do c_A and c_B behave according to the learning rule? In particular, what are the steady-state values of c_A and c_B? In this study, we have used a simple learning rule that assumes that, at the end of each trial, only synapses projecting to the chosen population undergo stochastic strengthening (if the choice is rewarded) or weakening (otherwise).

In general, the modification rates of potentiation (q_+) and depression (q_-) can be different. In the special case in which these two learning rates are equal, the steady state of synaptic strengths of the two sets of plastic synapses are approximately equal to the returns from the two choices [this approximation holds while the learning is slow (i.e., when q_+ and q_- are small)]. This can be shown by a simple calculation (Brunel et al., 1998). The probability of obtaining a reward on target $i = A$ or B, is equal to the number of rewards on that target divided by the total number of trials, which by definition is equal to the income from target i (I_i). If the probability of choosing target i is P_i then the average change in c_i in each trial is given by the following:

$$\Delta c_i = q_+(1 - c_i)I_i - q_- c_i(P_i - I_i).$$

The first term is the change attributable to potentiation in a rewarded trial (which occurs with the probability of I_i) and the second term is the change attributable to depression in a trial in which target i is chosen, but it is not accompanied by reward (which occurs with the probability of $P_i - I_i$). In the steady state, Δc_i should be zero which results in the following:

$$c_i^{ss} = \frac{q_+ I_i}{(q_+ - q_-)I_i + q_- P_i} = \frac{q_+ R_i}{(q_+ - q_-)R_i + q_-}, \qquad (7)$$

where $R_i = I_i/P_i$ is the return from choice i (i.e., the total number of reward obtained on choice i divided by the total number of choices for i).

It is thus clear that the steady state of c_i is a function of the return, R_i, from the choice i. In the special case in which $q_+ = q_-$, the steady state of c_i is equal to R_i. Even when q_+ and q_- are different, c_i is approximately a linear function of the return, $c_i^{ss} \simeq (q_+/q_-)R_i$, as long as $|q_+ - q_-|R_i$ is much smaller than q_-. The latter inequality generally holds when the return is significantly smaller than 1 (which is the case in the simulated experiment) and q_+ is not much larger than q_-.

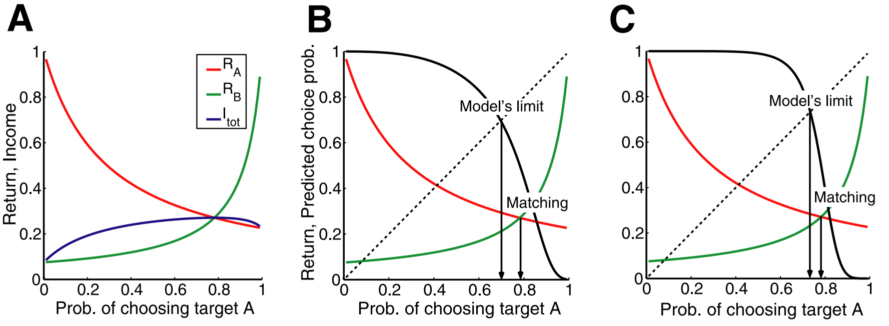

Figure 5. Mechanism of melioration and the limit for achieving perfect matching. ***A***, For a given choice probability, the return from each target (red for A and green for B) is computed and is plotted as a function of the probability of choosing target A (in all panels). The baiting probability on target A is three times the baiting probability on target B, and the overall baiting probability is equal to 0.3. Matching happens at a choice probability ($P_A = 0.782$) for which the returns from the two targets are equal. At this choice probability, the total income is optimal (blue curve) showing that, in this task, matching is optimal. ***B***, The choice probability for target A, predicted by the model (using Eq. 6) is shown in black for $\sigma = 10\%$. The steady state of the model is the point at which the predicted and current choice probabilities are equal, which is given by the intersection of the black curve with the diagonal line. The location of the steady state falls short of the prediction of matching, a phenomenon called undermatching. ***C***, For a smaller value of $\sigma = 5\%$, the steady state is closer to the prediction of matching (compare ***B***, ***C***).

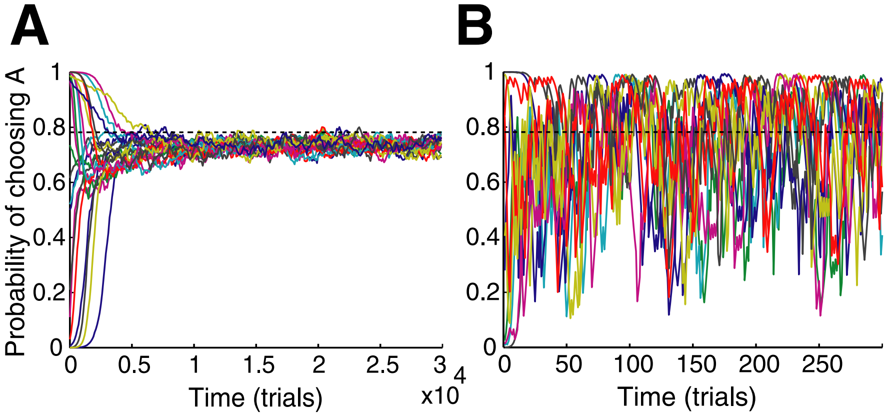

Figure 6. Time course of the choice behavior of the model in a matching task without COD penalty. The probability of choosing target A as function of time is plotted for different random initial values of c_A and c_B. The baiting probability ratio is similar to Figure 5 (3:1 in favor of target A). ***A***, The model reaches the steady-state choice behavior over a long time when the learning rates are very small ($q_+ = q_- = 0.0006$). ***B***, For more realistic values of learning rates ($q_+ = q_- = 0.06$), the model reaches the steady state quickly, but the fluctuations in the choice behavior are large. The dashed line shows the prediction by perfect matching ($P_A = 0.782$), and for both simulations σ is set to 5%.

Matching through probabilistic melioration

We have seen that, in our model, the return from each choice is represented in the synaptic strength of plastic synapses. The difference between synaptic strengths, $c_A - c_B$, determines the choice probability, which in turn modifies the returns. This interplay between the synaptic strengths (or equivalently returns) and the choice probability underlies dynamic decision of our model. Here, we show how this interaction can result in matching behavior. We shall first analyze an ideal situation to gain an intuitive understanding, and then consider more realistic simulations in the next subsection.

For simplicity, we focus on a discrete version of the concurrent variable-interval schedule (VI–VI), without imposing a change-over-delay penalty (see Materials and Methods). In the following, we shall begin by assuming that the model selects between the two targets with a given (current) choice probability. Based on the current value of choice probability, we then compute the returns (hence the synaptic strengths) and use Equation 6 to calculate the "predicted" choice probability from the model. We assess whether matching is achieved in the steady state when the predicted and current choice probabilities are equal (self-consistent). Specifically, for a given (current) choice probability, say for target A, P_A, the return from each target (R_A and R_B) can be computed easily (Heyman and Luce, 1979; Houston and Sumida, 1987). In Figure 5*A*, these returns are plotted (red and green curves) as a function of P_A, for given baiting probabilities. On the same plot, the total income from the two targets, $P_AR_A + P_BR_B = I_A + I_B = I_{tot}$, is plotted (blue curve).

When the returns from the two targets are equal, the total income is maximal (at $P_A = 0.782$). Therefore, in this task, matching corresponds to optimal behavior (Staddon and Motheral, 1978).

Now, in our model, the choice probability is a function of the difference between synaptic strengths, $c_A - c_B$ (Eq. 6). Assuming that the synaptic strengths are equal to the returns (red and green curves), the choice probability predicted by the model can be calculated according to Equation 6. This is plotted in Figure 5, *B* and *C* (black curve), for two different values of σ. At the intersection of the red and green curves ($P_A = 0.782$) where the returns from the two targets are equal, and hence $c_A = c_B$, the predicted choice probability by the model is equal to 0.5. If the return from target A is larger than the return from target B (for $P_A < 0.782$), the predicted choice probability is biased toward target A (>0.5) and when the return from target B is larger than the return from target A (for $P_A > 0.782$), the predicted choice probability is biased toward target B. In this sense, our model acts according to the melioration principle (Herrnstein and Vaughan, 1980; Vaughan, 1981; Herrnstein and Prelec, 1991), which states that the choice behavior should be biased toward the option with the higher return. However, decision is not deterministic in our model; the target with the higher return is chosen simply with some probability larger than 0.5.

Because of the probabilistic nature of our decision-making model, there is always a limit for approaching perfect matching. This concept is illustrated in Figure 5*B*. If the predicted choice probability computed by the model is greater than the current value of choice probability, P_A, then the model has a tendency to increase the probability of choosing target A. If the predicted choice probability is smaller than the current value of choice probability, then the model has a tendency to decrease the probability of choosing target A. The final state of the model (steady state) is the point at which the current and predicted choice probabilities are equal, that is, the intersection of the black curve and the diagonal line in Figure 5*B* ($P_A \approx 0.7$). This mechanism makes the model reach a choice probability that is generally smaller but close to that according to the matching law (0.782), a phenomenon called "undermatching." The extent of undermatching depends on the value of σ, so that, for a smaller value of σ, the steady state of the model is closer to the prediction of matching ($P_A \approx 0.73$ in Fig. 5*C* instead of $P_A \approx 0.7$ in Fig. 5*B*).

We found that this steady state of the model is stable, that is, the final state of the model is the same independent of the initial condition. Examples of the model choice behavior as a function of time are shown in Figure 6*A*, for which the learning rates are set to very small values. If the learning rates are large, the choice probability converges to a steady state fast, but fluctuations around the steady state are large (Fig. 6*B*).

So far, we have discussed the case in which the potentiation and depression rates are the same, so the synaptic strengths are equal to the returns. What happens if the rates of potentiation and depression are not equal? As we mentioned, if the overall baiting probability is small and q_+ is not much larger than q_-, c_A and c_B are still linear functions of returns from the two choices. For given R_i values, if $q_+ > q_-$, the differential input ($c_A - c_B$) is larger than the difference between returns (because the slope of c_i values as a function of return is >1). Because the choice probability P_A is a function of the ratio of the differential input to σ (Eq. 6), a larger differential input is equivalent to an effectively smaller value of σ, which results in better matching. If $q_+ < q_-$, the differential input is smaller than the difference in returns, which is equivalent to a larger value of σ; thus, the model shows more undermatching (data not shown).

Choice behavior of the model in a dynamic environment

In the previous subsection, we established that in a stationary environment for which the baiting probabilities stay constant, our model is able to reach a choice behavior close to matching. Now, we investigate whether this holds true in a dynamic environment, especially when the baiting probability ratio changes frequently between blocks of trials.

We simulate a matching task experiment in which the baiting probability ratio changes between blocks of 200 trials, similar to the task used by Sugrue et al. (2004a). An example of the choice behavior of the model in one simulated session of the experiment is shown in Figure 7*A*.

In each block of trials, the choice ratio, which is the slope of the cumulative choice plot, approximately matches the baiting probability ratio (Fig. 7*A*, black straight lines). Moreover, the instantaneous choice fraction closely follows the instantaneous reward fraction (Fig. 7*C*), an indication that matching is achieved dynamically in our model. The systematic trial-to-trial change in the choice behavior is determined by the ongoing changes in the synaptic strengths. To show these changes, synaptic strengths and choice probability are plotted during the same simulation (Fig. 7*D*). In each trial, only the synaptic strength that corresponds to the selected target undergoes small modification, and this modification is enough to alter the choice probability in each trial. In contrast, in each block of trials, the synaptic strengths fluctuate around different average values depending on the baiting probabilities in that block. In the example shown, these average values are the returns from each target. To demonstrate this point more clearly, we plot the average synaptic strengths in each block of trials versus the return from the corresponding choice in the same block (Fig. 8*A*).

Similar to the analytical prediction, if the two learning rates are equal, the averaged value of synaptic strengths in each block, $\langle c_A \rangle$ and $\langle c_B \rangle$, are close to the returns from choice A and B, respectively. In the cases in which the learning rates are not equal (Fig. 8*B*), average synaptic strength is still approximately a linear function of the return.

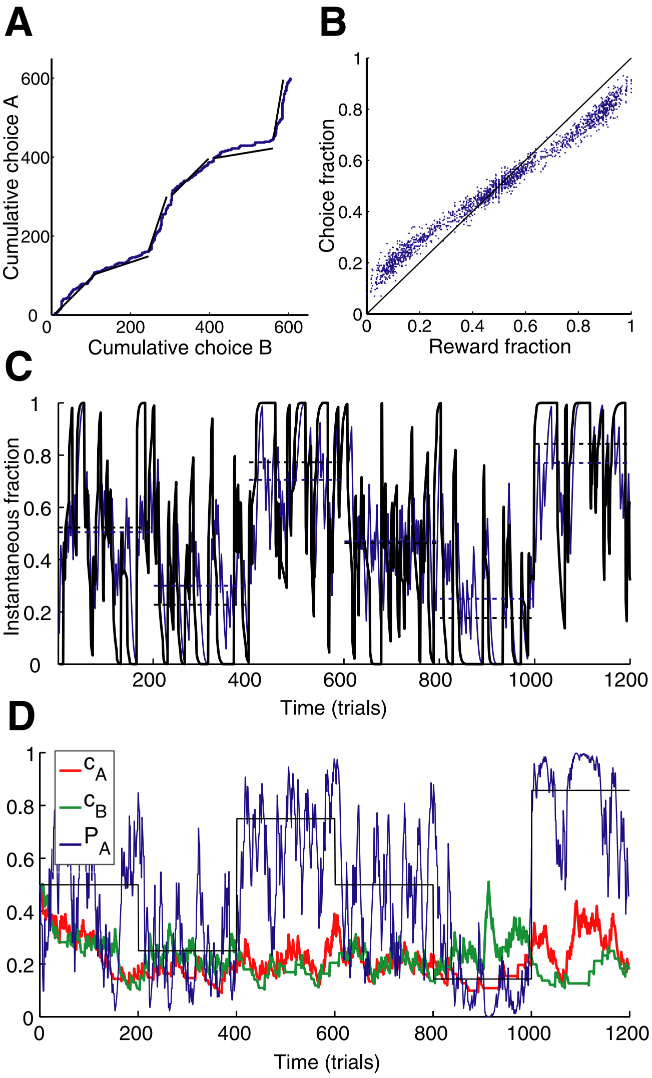

Figure 7. Model shows matching in a dynamic environment. ***A***, For one session of the simulated matching experiment, the cumulative choice on target A is plotted versus the cumulative choice on target B. The black straight lines show the baiting probability ratio in each block. The slope of the cumulative plot is equal to the choice ratio and is approximately equal to the baiting probability ratio. In this session the following baiting probability ratios are used in sequence [1:1, 1:3, 3:1, 1:1, 1:6, 6:1]. ***B***, Each point shows the blockwise choice fraction as a function of the blockwise reward fraction for a block of trials on which the baiting probabilities are held constant. The baiting probability ratios are selected from all possible ratios (see Materials and Methods). Most of the points fall close to the diagonal line (perfect matching), but the choice fraction is slightly lower than the reward fraction when the latter is larger than ½ (a phenomenon called undermatching). ***C***, The instantaneous choice (blue) and reward (black) fractions as a function of time computed for the same session shown in ***A***. The dashed lines show average choice and reward fractions for each block (in blue and black, respectively). To compute the instantaneous fractions, the choice and reward fraction are smoothed with a causal half-Gaussian filter with SD of six trials. The model is able to follow changes in the reward schedule. ***D***, The synaptic strengths, c_A (red) and c_B (green), and the choice probability (blue) as a function of time for the same data shown in ***A*** and ***C***. The thin black line indicates the baiting probability ratio in each block. In each block, the synaptic strengths fluctuate around the value of returns from the two choices. The model parameters for these simulations are $q_+ = 0.06$, $q_- = 0.06$, and $\sigma = 5\%$; and the length of each block is set to 200 trials.

To show the global choice behavior of the model, in Figure 7*B*, we plot the blockwise choice fraction (proportion of choice on target A in a block) versus the blockwise reward fraction (proportion of reward obtained from target A). The model shows good

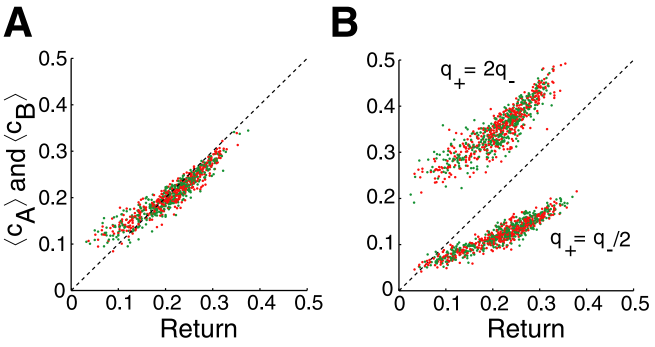

Figure 8. Plastic synapses approximately compute the return from each choice (or a linear function of it). Block-averaged synaptic strengths are plotted versus the obtained returns in the same block. Two colors represent two different choices (red for A; green for B). ***A***, The average synaptic strength is equal to the return from each choice, when the two learning rates are equal ($q_+ = q_- = 0.06$). ***B***, If the potentiation rate is greater than the depression rate ($q_+ = 0.06$; $q_- = 0.03$), the average synaptic strength is larger than the return (top points), whereas, if the depression rate is greater than the potentiation rate ($q_+ = 0.03$; $q_- = 0.06$), then the synaptic strength is smaller than the return (bottom points). These data points are obtained from 15 simulated sessions of the matching task in which all possible baiting probability ratios are used (see Materials and Methods), and the length of each block is set to 200 trials. For all simulations, σ is set to 5%.

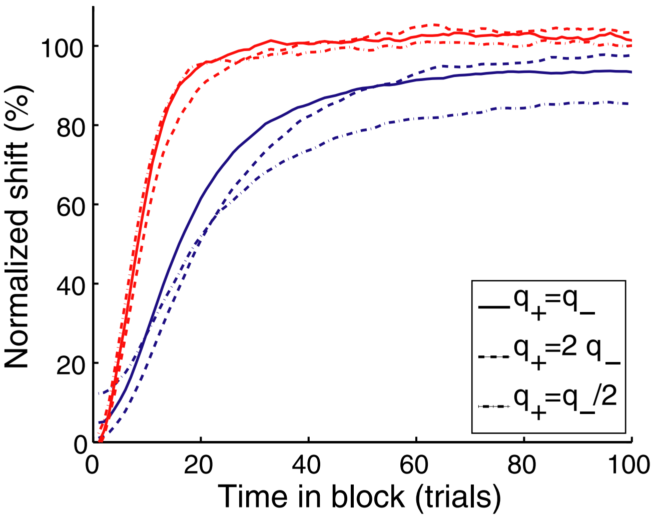

Figure 9. Adaptation to change in the reward schedule. The average time course of adjustment for the choice (blue) and reward (red) fractions in a new block of trials are plotted for three different values of learning rates: $q_+ = q_- = 0.06$ (solid curves); $q_+ = 0.06$, $q_- = 0.03$ (dashed curves); and $q_+ = 0.03$, $q_- = 0.06$ (dot-dashed curves). The choice and reward fractions are normalized, so 0% shift indicates the same fraction as the fractional baiting probability in the previous block, and 100% shift indicates the same fraction as the fractional baiting probability in the current block. The instantaneous choice and reward fractions are computed using a causal half-Gaussian filter (SD = 6 trials). The shift in the reward fractions happens in the span of 10–15 trials, and it is approximately independent of the learning rates. The shift in the choice fraction reaches an asymptotic value after 30–40 trials, which is dependent on the learning rates. If $q_+ > q_-$, the shift in choice behavior is slower right after the block transition, but its reaches a higher value later on. These results are obtained from 500 simulated sessions of matching task with all possible baiting probability ratios (see Materials and Methods). The length of each block is set to 200 trials, and σ is set to 5% in all simulations.

matching, that is, the choice fraction in each block of trials is approximately equal to the reward fraction in that block. Moreover, in blocks in which target A is richer (reward fraction >0.5) the choice fraction is usually smaller than the reward fraction. This general tendency of the model, called undermatching, has also been observed in matching task experiments in monkeys (Anderson et al., 2002; Sugrue et al., 2004a; Lau and Glimcher, 2005b).

In the previous section, we showed that, in a stationary environment, the probabilistic nature of decision making in our model results in undermatching. In a dynamic environment, undermatching becomes even more prominent, because after a change in the reward schedule, it takes a few trials for the choice behavior to be shifted according to the new reward schedule. To illustrate how fast the model is able to shift its choice behavior between blocks of trials, following the same method (with a slight modification) used by Corrado et al. (2005), we plot the normalized shift (which is the shift per trial normalized by the programmed reward fraction shift) in choice and reward fractions after each block transition (Fig. 9).

Similar to the monkey experiment, it takes ~30–40 trials for the model to completely shift its choice behavior after a block transition (Corrado et al., 2005). Furthermore, the relative values of the learning rates affect how quickly shifts can take place. For example, if $q_+ > q_-$, the shift is slower at the beginning of a block transition but reaches a higher asymptotic value later on. This happens because in this case the difference between the average value of synaptic strengths is large, and with a slow depression rate it takes more time to reverse this difference in the new block (in which the more rewarding choice is different from the last block).

Stay length and switch probability

A few behavioral studies of matching tasks in pigeons and rats have shown that switching between choices is approximately a stochastic process that only depends on the reinforcement rate on those choices (Heyman and Luce, 1979; Gallistel et al., 2001). This means that the probability that the animal switches from one choice to the other is almost independent of the time that it has spent on the current choice. Based on these results, it has been claimed that matching cannot be generated by a mechanism that involves feedback (Heyman, 1979; Gallistel et al., 2001).

We assessed the statistics of choice behavior separately for each block of trials with different baiting probability ratios. Note that for all analyses presented in this section, the mandatory movements after switches are removed because in these trials the choice behavior follows a deterministic rule. In Figure 10*A*, the distribution of stay lengths (number of consecutive choices on one target) is plotted for the two targets in different blocks of the experiment. In addition, each stay length histogram is fitted with an exponential distribution (black curves).

Consistent with the experimental observation (Corrado et al., 2005), the distribution of stay lengths on each target is approximately exponential, although a small deviation can be seen clearly. Moreover, for the target with a larger baiting probability the stay length distribution has a larger mean.

If switching between the two targets is a completely stochastic process with a rate determined by the baiting probability, one expects that probability of staying longer than a given stay length (survival probability) is an exponentially decreasing function of the stay length. As shown in Figure 10*B*, the probability of staying longer than a given stay length is approximately a linear function of the stay length in a semilog plot. The negative of the slope in this plot is approximately equal to the probability of switching to the other target (Gallistel et al., 2001). We also have computed the probability of switching as a function of the stay length (Fig. 10*C*). If the survival probability is perfectly monoexponential, then the switching probability should be independent of the stay length. We found that this is only approximately true, for long stays.

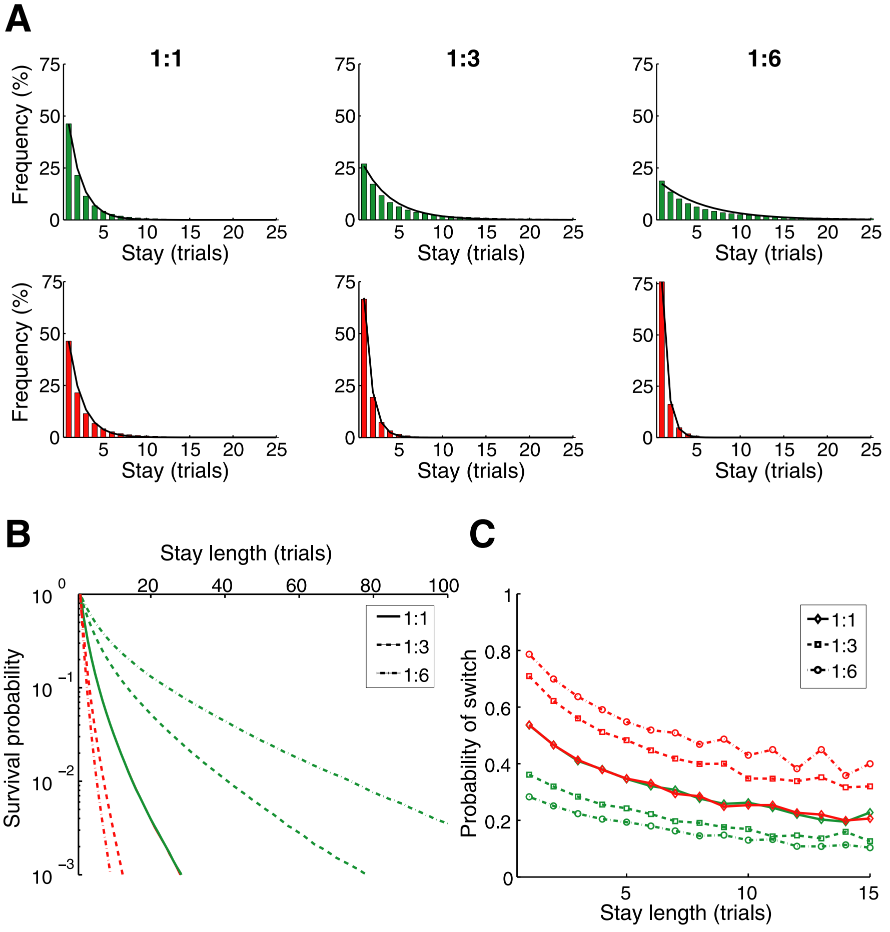

Figure 10. Statistics of the stay length and switch probability. ***A***, The distribution of stay lengths in each block (with fixed baiting probabilities) is approximately an exponential and depends on the baiting probability. The baiting probability ratio in each block is reported on each plot (in the favor of target B). For the 1:1, 1:3, and 1:6 schedule, the mean stay length for target A (red histograms) is 2.65, 1.63, 1.38, and for target B (green histograms) is 2.65, 5.71, 9.66, respectively. The black curves show the fitting with an exponential distribution. ***B***, Log cumulative probability of staying longer than a given stay length (survival probability) is plotted separately for different targets in blocks of trials with different baiting probability ratios. Red and green curves, The survival probabilities for the targets A and B, respectively. The baiting probability ratio for each block is shown in the inset. The negative of the slope in this plot is equal to the probability of switching to the other target. ***C***, The probability of switching as a function of the stay length. These statistics are obtained from 5000 simulated sessions in which the baiting probability ratio is [1:1, 1:3, 1:6] and the overall baiting probability is fixed to 0.3 (block length is equal to 2000 trials). The model parameters are $q_+ = q_- = 0.06$ and $\sigma = 5\%$.

Indeed, the probability of switching is a decreasing function of the stay length, and it reaches a steady state after ~10–15 trials. Note that the exact value of switch probability and its steady state depends on the model parameters, but its qualitative form is the same. Furthermore, the probability of switching is a function of baiting probability, so it is larger for the target with a smaller baiting probability (Fig. 10*C*).

In our model, staying on a choice or switching to another choice is a stochastic process with a probability that is determined by the state of plastic synapses in each trial (except for the mandatory movement after a switch). Note that the probability that the model stays on a target for a large number of trials is very small (especially when the baiting probability is low), and because of the stochastic nature of our model a long stay requires harvesting of a few rewards on that target. As a result, the probability of switching after a few stays decreases as a function of the stay length. Our results demonstrate that in a model based on feedback, like our model, the probability of switching is determined by the baiting probability. In addition, the slow change of the probability of switching as a function of the stay length may not be incompatible with the experimental observation of approximately constant switch probability.

Robustness of the model

Although in the previous simulations a specific set of parameters is used, matching behavior can be achieved over a wide range of parameters. The behavior of the model is quantified, using a sequence of blocks of trials with different baiting probability ratios, when the learning rates, q_+ and q_-, are varied in a broad range (with fixed σ for the noise level). We quantify the performance of the model by the ratio of the average reward rate to the overall baiting probability, as in the monkey experiment. The performance of the model is assessed with a sequence of blocks with different baiting probability ratios (see Materials and Methods) and is plotted in Figure 11*A*. For most of the potentiation and depression rates, the performance of the model is high compared with the monkeys' performance [~72% in the study by Sugrue et al. (2004a)] and does not vary significantly over the range of the parameters.

The performance is relatively low in two cases. In the first case, the potentiation rate, q_+, is much larger than the depression rate, q_- (for moderate values of q_-). This condition results in long stays on the richer target in each block (see the switching probability in Fig. 11*C*), so the model is slow in shifting its behavior between blocks of trials and loses some of the rewards. In the second case, the depression rate is much larger than the potentiation rate so any unrewarded trial gives rise to a strong reset of the synaptic strength for the chosen target. As a result, the model alternates frequently between the two choices (Fig. 11*C*). For small values of q_-, the performance of the model is high, because in these cases both synaptic strengths saturate and reach a value close to 1, and as a result the choice behavior becomes more random.

We define the "deviation from matching" as the average absolute difference between choice and reward fractions on each block. This quantity is more strongly dependent on the learning rates, although for most of the parameter space its value is small (Fig. 11*B*). Similar to the results for the choice behavior in a stationary environment, generally better matching can be achieved with a potentiation rate larger than the depression rate.

As we mention for this task when the baiting probabilities are constant, matching is optimal. However, if the baiting probability changes between blocks of trials, this statement no longer holds. In a stationary environment, good matching can be achieved when the potentiation rate is higher than the depression rate and there are not many switches between the two choices. But with these conditions, in a dynamic environment the choice behavior cannot be shifted quickly, and as a result the performance will deteriorate. So in a dynamic environment, the model parameters that result in the best matching behavior do not correspond to

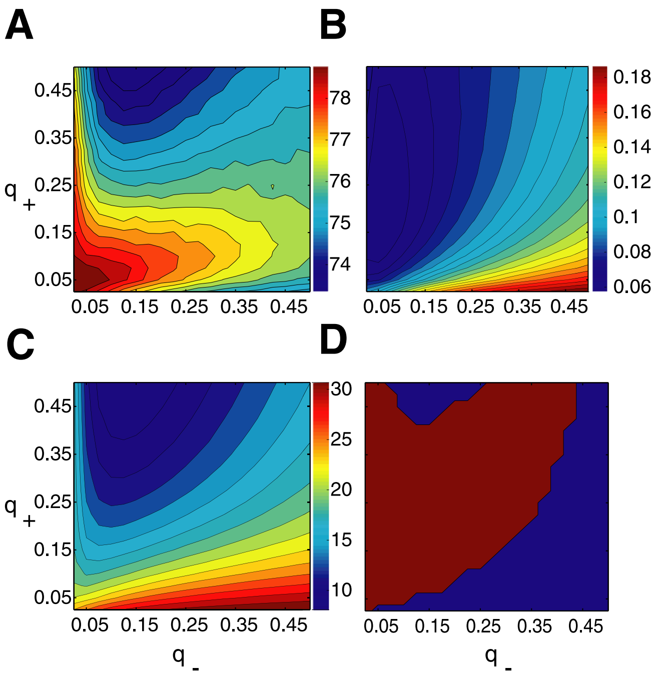

Figure 11. Model shows matching behavior over a wide range of parameters. ***A***, The performance of the model, defined as the ratio of the average reward rate to the overall baiting probability I_{tot}/λ_{tot}, is generally high and does not change significantly with learning rates (only a few percent change). ***B***, The "deviation from matching," computed as the average of absolute difference between choice and reward fractions on each block, is small over a wide range of learning rates. This indicates that, for a wide range of learning rates, a choice behavior close to matching can be achieved. ***C***, The switching probability (expressed in percentage), the total number of switches between the two choices divided by the total number of trials, is strongly dependent on the learning rates. For large values of q_-, the switching probability is high, but a large value of q_+ reduces the switching probability. ***D***, The range of parameters for which the model shows an adequate matching behavior is plotted in red, that is, when $I_{tot}/\lambda_{tot} > 0.74$ [this quantity is ~0.72 for the monkeys (Sugrue et al., 2004a)] and the deviation from matching is <0.1. For each set of model parameters, all average values are obtained from 1000 simulated sessions of the experiment (see Materials and Methods). The length of each block is set to 200 trials and $\sigma = 5\%$.

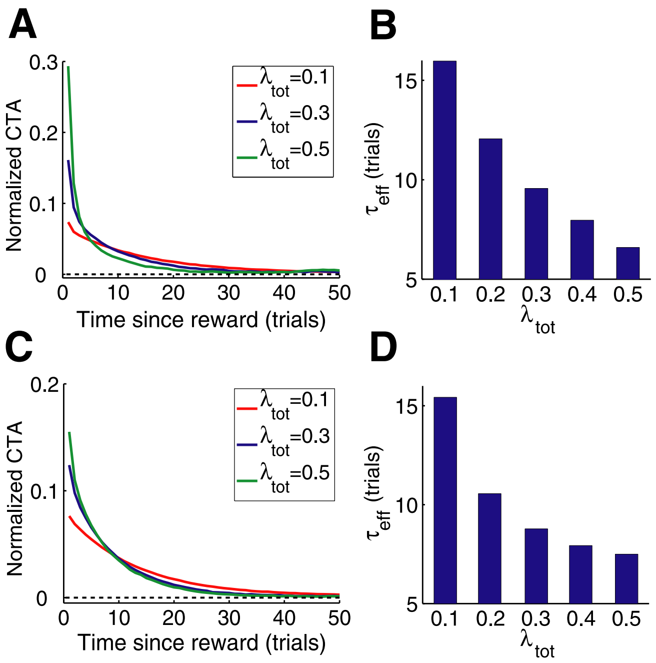

Figure 12. Time integration of past rewards and its dependence on the overall baiting probability. ***A***, The choice-triggered average of rewards, extracted from the model choice behavior, is plotted for three different values of overall baiting probability (for a fixed set of the model parameters; $q_+ = 0.15$, $q_- = 0.05$, $\sigma = 25\%$). As the overall baiting probability, λ_{tot}, increases, recent rewards have a stronger effect, but the effect of past rewards decays more quickly. ***B***, The effective time constant of CTA, defined as the weighted sum of the two extracted time constants, is plotted as a function of the overall baiting probability. As the total reward rate increases, the effective time constant decreases. ***C***, For a different set of model parameters ($q_+ = 0.06$; $q_- = 0.06$; $\sigma = 5\%$), CTA extracted from the model choice behavior is plotted for three different values of overall baiting probability. ***D***, The effective time constant of CTA extracted from the choice behavior of the model for the same set of parameters as in ***C***.

optimal performance. There is an intermediate range of learning rates that results in a large reward rate and also reasonable matching behavior. We define a set of the model parameters as suitable, if for such a set of parameters the performance of the model is >74% and the deviation from matching is <0.1. These sets of parameters are shown in Figure 11*D* in red. Note that suitable behavior can be achieved for a wide range of the learning rates, so that in order for the model to perform the matching task, fine-tuning of the learning parameters is not necessary.

Dependence of the choice on the past rewards

To quantify the dependence of choice in each trial on the past history of reward, we follow Sugrue et al. (2004a) (Corrado et al., 2005) to calculate what they termed as the "choice-triggered-average of rewards (CTA)." This quantity measures how choice in the current trial is influenced by the harvested rewards in the past trials. Here, we are mainly interested in how the form of CTA, extracted from the model choice behavior, is influenced by external factors such as the length of blocks in which the baiting reward rate is constant, and the overall baiting probability in the experiment.

In general, the form of CTA in our model can vary, depending on the learning rates and the noise level σ. Interestingly, we find that the shape of CTA is independent of the length of the blocks, consistent with the observation of Sugrue et al. (2004a) that CTA is the same no matter which part of the block is used for analysis. To compare our results with the Sugrue et al. (2004a) experiment, we choose a set of the model parameters that results in a CTA close to the CTA extracted from one of the monkeys in their experiment. Using the same parameters, we allow the model to play the same task with different overall baiting probabilities (λ_{tot}).

The CTA for three different values of overall baiting probabilities is shown in Figure 12*A*. When the overall baiting probability is high, the dependence of the choice on the recent harvested rewards is stronger, but this dependence decreases more rapidly with time (in unit of trial).

In contrast, when the overall baiting probability is low, the choice is influenced by past rewards over a longer time. To quantify this dependence, the CTA obtained from the behavior of the model is fitted with the sum of two exponentials as follows:

$$\mathrm{CTA}(t) = \omega N_s e^{-\frac{t}{\tau_s}} + (1-\omega) N_l e^{-\frac{t}{\tau_l}}, \qquad (8)$$

where t is the trial number, N_s and N_l are the normalization factors for each exponential, ω is the weighting factor, and τ_s and τ_l are the short and long time constants, respectively. The sum of two exponentials provides a good fit for the CTA extracted from the monkeys' data (Corrado et al., 2005). The results of fitting show that, when the overall baiting probability increases, the longer time constant, τ_l, and its relative weight, $1 - \omega$, decrease (data not shown). To simplify the comparison, we define an effective time constant, $\tau_{eff} = \omega\tau_s + (1 - \omega)\tau_l$. The effective time constant displays the approximate timescale over which the integration of past rewards is performed. As shown in Figure 12*B*, the effective time constant decreases as the overall baiting probability increases.

For the set of model parameters used in most of the simulations ($q_+ = 0.06$; $q_- = 0.06$; $\sigma = 5\%$), the extracted form of CTA is different from the experimental observation (Fig. 12*C*,*D*), namely, in this case the relative weight of the shorter time constant is small. Additional study shows that generally a more biexponential CTA can be achieved if the choice behavior of the model is more random (larger value of σ), which results in more undermatching. Although generally the quality of matching in monkeys is poorer than the model, these results may indicate that there are other factors that influence the monkey's choice behavior, such as past choices of the monkey, which are not included in the present model.

Discussion

Recent neurophysiological studies of nonhuman primates performing probabilistic decision making tasks showed that single-cell activities in certain brain areas are modulated by the subjective values of choice options (Kawagoe et al., 1998; Leon and Shadlen, 1999; Platt and Glimcher, 1999; Lauwereyns et al., 2002; Montague and Berns, 2002; Shidara and Richmond, 2002; Barraclough et al., 2004; Sugrue et al., 2004a, 2005; Samejima et al., 2005; Hikosaka et al., 2006). In this study, we addressed the question of how these subjective values may be computed mechanistically and used to generate choice behavior. We showed that plastic synapses that undergo stochastic reward-dependent modification can integrate past rewards within a finite time window. This is because synapses have a limited number of discrete states, so reward history in the remote past is forgotten, resulting in a finite integration time. The strengths of these synapses influence the choice behavior in any given trial, and in turn they are modified depending on the choice made and the resulting outcome in that trial. This two-way interplay between synaptic plasticity and the decision process gives rise to trial-by-trial adaptive choice behavior in a dynamic and stochastic environment. In this work, we applied our model to a matching task paradigm and showed that, under certain conditions, the average strength of plastic synapses onto a choice-selective neural population is equal to the return from that choice. Therefore, we propose that subjective values can be computed dynamically, in the form of return, at the synaptic level. Furthermore, decision neurons that receive inputs from these plastic synapses exhibit graded levels of activity. The latter is modulated by the choice of the network and the difference in the average synaptic inputs to the competing neural populations. In this way, the subjective values computed at a synaptic level become observable in the spike firing neural activities of a decision-making network.

Learning rule

Learning that depends on reinforcement feedback signals (Sutton and Barto, 1998) is believed to underlie many adaptive behaviors in a natural environment. Evidence suggests that dopamine in the brain acts as a common currency for a reward signal (Schultz, 2000, 2006), and modulates synaptic plasticity (Jay, 2003). For instance, at the corticostriatal synapses onto the medium spiny projection neurons in striatum, long-term plasticity depends on stimulations of dopamine neurons (Reynolds et al., 2001; Reynolds and Wickens, 2002) (but see Fino et al., 2005). Based on these observations, Reynolds and Wickens (2002) proposed a three-factor synaptic plasticity rule, in which synaptic modifications depend on presynaptic and postsynaptic neural activities as well as a dopamine signal. Other studies indicate that, in the rat prefrontal cortex, the induction of long-term depression and long-term potentiation at glutamatergic synapses is modulated by dopamine (Otani et al., 2003), and that D_1 receptors play a key role in such bidirectional modulation of plasticity (Huang et al., 2004).

In this work, we sought to implement such a three-factor learning rule in a biophysically plausible manner (Fusi et al., 2005a). Our learning rule is Hebbian and depends on the coactivation of presynaptic and postsynaptic neurons. Moreover, individual synapses have two discrete states (Petersen et al., 1998; O'Connor et al., 2005), and plasticity occurs as a stochastic process (Amit and Fusi, 1994; Fusi, 2002). The fact that synapses are bounded is important, because in this way the number of available synaptic states are limited and this enables the model to forget the past outcomes naturally. The stochastic nature of plasticity implies that modifications occur at every single trial, which form the basis of adaptive decision process; yet the average synaptic changes take place over many trials (determined by the learning rates), over a timescale compatible with that of experimental protocols for the induction of long-term synaptic potentiation or depression (Malenka and Nicoll, 1999; Bi and Poo, 2001). Finally, the direction of modification (potentiation versus depression) is reversed by the presence/absence of reward (see below for variants of this learning rule). Seung (2003) also considered a reward-dependent synaptic plasticity rule and proposed an algorithm based on a reward signal that modulates the probability of stochastic release of transmitters. By design, Seung's algorithm maximizes rewards under general conditions, whereas this is not guaranteed with our model. In contrast, Seung's model performs better if an all-or-none reward signal, like dopamine, is delivered every time a presynaptic spike is fired; how this may be accomplished biologically remains unclear. Delivery of reward signal with a delay, requires an integration of eligibility trace over a timescale of seconds during the decision process and makes the learning process very slow. In contrast, in our model, plasticity occurs only once in a trial, rather than continuously, at the time of potential reward delivery.

We focused on a specific learning rule in which plasticity takes place only when the postsynaptic neurons fire spikes at a high rate, in other words, only for plastic synapses projecting to the neural population that has won the competition in a given trial. As we have shown, our model exhibits satisfactory matching behavior comparable with observations in the experiment of Sugrue et al. (2004a), robustly for a wide range of learning rates for potentiation and depression, respectively. The only condition is that the potentiation rate should be similar to or larger than the depression rate. We also explored other variants of learning rules. For example, plasticity could take place without requiring high firing activity of postsynaptic neurons. Thus, in each trial both sets of plastic synapses onto the two competing neural populations are modified. In this situation, the synaptic strength is not a function of return. Instead, it is a function of income (for a more detailed description of such a model and its behavior, see supplemental material, available at www.jneurosci.org).

Moreover, if the decision criterion is given by the fractional income [i.e., $P_A = I_A/(I_A + I_B)$ (like in the model of Sugrue et al., 2004a)] instead of a sigmoid function of the differential income, the choice behavior may become unstable. The instability happens in the income-based model for the following reason: if one of the targets is consecutively chosen, although few rewards are obtained, the income on the chosen target fluctuates around some level while the income on the unchosen target goes to zero. This further decreases the probability of selecting the unchosen target and results in repeated selection of one of the targets. In contrast, the return changes for the selected target only, hence such instability does not occur in a return-based decision model.

If the rule is such that, in a rewarded trial, not only synapses are strengthened with high postsynaptic activity (of neurons selective to the chosen target), but also weakened with low postsynaptic activity (of neurons selective for the unchosen target), then the model tends to select the target with a higher baiting probability excessively, a phenomenon called "overmatching." One could argue that overmatching may be avoided using a learning rule according to which, in an unrewarded trial, potentiation occurs with low postsynaptic firing rates (for those synapses projecting to neurons selective for the unchosen target). This plasticity rule seems biophysically implausible. Moreover, this rule typically leads to more undermatching than the rule presented in this paper does, and is thus functionally undesirable (large deviation from matching). The general conclusion is that the most suitable and robust learning rule for the matching task is the one in which only plastic synapses related to the selected choice undergo plasticity, and this rule is qualitatively in agreement with the available experimental evidence for reward-dependent synaptic plasticity.

Mechanisms of matching behavior

Although matching behavior has been observed in many different experimental paradigms, how it is achieved by a local (trial-to-trial) decision process is still not fully understood. In one scenario, Gallistel et al. (2001) proposed that a local mechanism based on ideal detectors of changes in reward rates can account for matching behavior. Another proposal relies on the idea that matching is a manifestation of reward maximization (Staddon and Motheral, 1978; Baum, 1981). A third theory, called "melioration" (Herrnstein and Vaughan, 1980; Williams, 1988; Herrnstein and Prelec, 1991), posits a decision dynamics in which the subject chooses the behavioral alternative that provides the higher local reinforcement rate (or return) at that time. In the special case of the concurrent variable-interval schedule, this local mechanism produces global matching behavior, because an increase (decrease) in the selection of one option decreases (increases) the return from that option. Therefore, an equilibrium is reached when the returns from the two alternatives are equal. However, in general, melioration can result in a behavior different from matching. In fact, in experiments in which melioration, matching, and maximization give different predictions, behavioral data were consistent with the melioration theory (Vaughan, 1981).

Two key issues have been left unresolved in the melioration theory. First, as stated by Williams (1988): "the most fundamental problem faced by melioration theory is the specification of the method by which local reinforcement rates are calculated." Second, melioration has often been taken to mean "choose the option with the highest return among all possible alternatives." Although this deterministic rule yields matching as the steady state, the stability of that behavior is not guaranteed. The neural model reported in this paper sheds insights into both issues.

Our model proposes a neurobiological implementation of melioration. First, in our model, a local estimate of return (or a function of return) on each choice is computed by synapses that undergo reward-gated stochastic plasticity. We found that there is a tradeoff between the accuracy of the estimated return and the flexibility of choice behavior. If the learning rates are low, synapses can integrate rewards over a long period of time and the estimation of return would be accurate. However, this means that the system cannot adapt quickly when the reward schedule changes frequently in an uncertain environment. In contrast, if the learning rates are higher, the behavior is more flexible, but the reward integration is more local in time and the estimated return is noisier. This raises the interesting question of whether learning rates themselves should be plastic (meta-learning) and adjustable according to behavioral demands (Doya, 2002; Schweighofer and Doya, 2003). We intend to address this question elsewhere.

Second, in our model, decision making is not deterministic even if the returns of options are known. Instead, we showed that, in a recurrent circuit of spiking neurons, the choice probability is a softmax (sigmoid) function of the difference in the returns (coded by the strengths of synapses to the two competing neural populations). This stochasticity is attributable to irregular spike discharges, a characteristic feature of cortical neurons (Softky and Koch, 1993; Shadlen and Newsome, 1994; van Vreeswijk and Sompolinsky, 1996; Compte et al., 2003). The more variable the neuronal firing activity, the less steep is the softmax function (with a larger σ). Thus, too much noise would mean a very graded softmax decision criterion; the choice behavior would be essentially random and far from matching. In contrast, with negligible noise (small σ value), the system has a tendency to only choose the target with a higher return, and this may result in instability of the choice behavior. To avoid this kind of instability, the network should be able to make decisions probabilistically. We also showed that probabilistic decision making imposes a limit on how close matching can be achieved. Therefore, our model provides a possible explanation, in terms of neural network constraints, for undermatching, a phenomenon widely observed across different species (Baum, 1974, 1979; Davison and Baum, 2000; Anderson et al., 2002; Sugrue et al., 2004a; Lau and Glimcher, 2005b).

Our model has similarities to, as well as differences with, other recently proposed models for matching behavior. For example, the model of Sugrue et al. (2004a) assumes that local incomes on two options are computed by a leaky integrator and then these quantities are used to compute the local fractional income. If the instantaneous probability of choice is equal to the local fractional income, the model obeys the matching law locally. This model provides a good account of monkeys' behavioral data but leaves open mechanistic questions, such as how integration over the income is done, how the time constant for the integration is determined in the circuit, how local fractional income is calculated (which requires two additional computations, addition and division), and how it can be translated to choice probability. In a revision of this model, Corrado et al. (2005) replaced the decision rule according to fractional income, by a softmax function of the difference in local incomes. Our model represents a biophysical instantiation of that scenario, except that it uses return rather than income.

Furthermore, our work suggests a synaptic mechanism for the linear filter (called CTA) that has been deduced from behavioral data and hypothesized to subserve reward integration (Sugrue et al., 2004a; Corrado et al., 2005; Lau and Glimcher, 2005b). In our model, the time constants of CTA are determined by the model parameters. Hence, the experimentally observed CTAs can be associated with biophysical quantities like potentiation and depression rates at the synaptic level. Importantly, we showed that the overall baiting probability influences the form of CTA, so that the integration times are stretched or contracted depending on how abundant rewards are in the environment. Specifically, we showed that, when the overall baiting probability is lower (higher), because synaptic modifications depend on reward frequency, the effective integration time becomes larger (smaller) so that rewards are integrated over a longer (shorter) timescale, which makes sense functionally. This prediction of our model can be readily tested by varying the overall reward rate in matching task experiments. Finally, our model semiquantitatively reproduces electrophysiological data recorded from behaving monkeys, whereas previous models (Sugrue et al., 2004a; Corrado et al., 2005; Lau and Glimcher, 2005b) were mostly focused on behavioral data.

LIP neurons: representation of decision or value?

Neurons in the LIP area of the posterior parietal cortex show activity that is believed to be important for guiding saccadic eye movements. In experimental studies of the matching task, it has been shown that these neurons carry information about the impending movements and the subjective value of those movements. Platt and Glimcher (1999) showed that the activity of some LIP neurons is modulated by the gain of the choice into the response field (RF) of the neuron. Sugrue et al. (2004a) showed that activity of LIP neurons is modulated by the impending choice and the fractional income for that choice. Importantly in a given trial, the activity of a neuron is higher if the monkey's choice is into the RF of the neuron than if the monkey's choice is out of the RF of the neuron. In addition, for a fixed choice, the activity of a neuron is higher when the local fractional income of the RF target of the neuron is higher. In a later paper, Corrado et al. (2005) showed that, in fact, their neural data were better described as being correlated with the difference between the local incomes from the two targets, rather than with the fractional income.

Although the graded activity of neurons in area LIP carries information about the value for each choice, there is evidence that the valuation is not computed in LIP. Indeed, LIP neurons are spatially selective and not color selective, whereas in this task the rewarding value of each target is coded by the target color and not by its location. Moreover, the time course of neural activity becomes differentiated according to the income level of a given target, at least 100–200 ms after the stimulus onset, indicating that the value-related information originates from somewhere else.

Our working hypothesis is that the primary role of LIP neurons is to generate a decision about saccadic eye movement, based on an integration of two types of inputs: spatial target and its rewarding value. According to this view, for a given neuron (and a selected target), the firing activity depends parametrically on its overall input. Thus, graded activity emerges whenever the trials are sorted according to different levels of the overall input of the neuron, regardless of whether it is a sensory or reward signal, or a combination of both. This interpretation is consistent with the observation that graded LIP neural activities, similar to those in the study by Sugrue et al. (2004a), were also found in a visual motion direction discrimination task in which the differentiating factor is motion coherence (sensory information) rather than rewarding value (Shadlen and Newsome, 2001; Roitman and Shadlen, 2002). It is also in line with the finding that, when the two choices about motion direction are associated with different amounts of reward, the subject's psychometric function is shifted in such a way as if reward magnitude provided an extra signal that is additive to the sensory information about the motion direction (Rorie and Newsome, 2004).

Here, we showed that a model based on this idea reproduces graded neural activities observed in LIP. In our model, neurons are responsible for making decisions, hence the spiking activities are correlated with and give rise to choices. In addition, because input synapses to these neurons encode reward history, neuronal firing rates naturally reflect the target rewarding values. Because the choice probability and returns are directly related to each other (via the softmax function), it is impossible to dissociate the two. Similarly, in a physiological experiment, correlations between neural activity and the subjective value of choice options do not necessarily imply that the recorded neurons (like LIP cells) are primarily involved with valuation rather than decision making.

It is worth noting that, as our results here demonstrate, graded activities are compatible with the attractor dynamics of our model. Indeed, although an attractor network displays stable activity states in the absence of direct stimulation (e.g., during a delay period of working memory), it is readily configurable by external inputs and can depend on input strength in a graded manner. Moreover, the other aspect of the observed graded LIP activity, namely the divergence over time of firing activities corresponding to two alternative choices, is explained in our model by effective mutual inhibition between the two selective populations. Because of this competition, when the firing rate of one selective population is high, that of the other selective population goes down. This is similar to what has been observed in the experiment of Sugrue et al. (2004a). In their experiment, if the left choice has the highest local fractional income and it is selected, then neurons with RF on the left target have the highest level of activity and neurons with RF on the right target have the lowest level of activity. If we assume that neurons with RF on the left and right target belong to two different pools of neurons in LIP, then the most plausible explanation for the above observation is the existence of competition between these two pools of neurons. This, again, is consistent with the suggestion that the LIP neurons behave like decision makers, or have an important role in the decision-making processes.

In this paper, we focus on a microcircuit model endowed with synaptic plasticity. In all likelihood, this model will need to be expanded, and the following alternative scenarios should be considered in future studies. First, if integration of past rewards is performed by plastic synapses, it is an open question as to the precise locus (or loci) of such plasticity. In addition to LIP, other candidate sites include corticostriatal synapses in basal ganglia, or synaptic connections in the orbitofrontal cortex (Schultz, 2000). For example, a new study showed that postsaccadic activity of caudate neurons encodes the preceding saccade and/or reward delivery in a matching task experiment (Lau and Glimcher, 2005a). Therefore, a large-scale network with multiple interacting brain areas should be investigated. Secondly, it is conceivable that past rewards can be integrated by cellular mechanisms in single neurons, instead of plastic synapses. There is evidence that neural activity in the dorsolateral prefrontal cortex (Barraclough et al., 2004) and orbitofrontal cortex (Sugrue et al., 2004b) is modulated by reward signals across trials. However, it remains an open question whether these reward-modulated neural activities are generated by a cellular or synaptic mechanism. Additional experimental and computational work will shed light on this fundamental question about the neurobiological basis of choice behavior.

References

Amit DJ, Fusi S (1994) Dynamic learning in neural networks with material synapses. Neural Comput 6:957–982.

Anderson KG, Velkey AJ, Woolverton WL (2002) The generalized matching law as a predictor of choice between cocaine and food in rhesus monkeys. Psychopharmacology (Berl) 163:319–326.

Barraclough DJ, Conroy ML, Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci 7:404–410.

Baum WM (1974) 2 Types of deviation from matching law-bias and undermatching. J Exp Anal Behav 22:231–242.

Baum WM (1979) Matching, undermatching, and overmatching in studies of choice. J Exp Anal Behav 32:269–281.

Baum WM (1981) Optimization and the matching law as accounts of instrumental behavior. J Exp Anal Behav 36:387–403.

Bi G, Poo M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev Neurosci 24:139–166.

Brunel N, Wang X-J (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J Comput Neurosci 11:63–85.

Brunel N, Carusi F, Fusi S (1998) Slow stochastic hebbian learning of classes of stimuli in a recurrent neural network. Network 9:123–152.

Compte A, Constantinidis C, Tegner J, Raghavachari S, Chafee MV, Goldman-Rakic PS, Wang X-J (2003) Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. J Neurophysiol 90:3441–3454.

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

Davison M, Baum WM (2000) Choice in a variable environment: every reinforcer counts. J Exp Anal Behav 74:1–24.

de Villiers PA, Herrnstein RJ (1976) Toward a law of response strength. Psychol Bull 83:1131–1153.

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

Doya K (2002) Metalearning and neuromodulation. Neural Netw 15:495–506.

Fino E, Glowinski J, Venance L (2005) Bidirectional activity-dependent plasticity at corticostriatal synapses. J Neurosci 25:11279–11287.

Fusi S (2002) Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. Biol Cybern 87:459–470.

Fusi S, Asaad WF, Miller EK, Wang X-J (2005a) A microcircuit model of arbitrary sensori-motor mapping: learning and forgetting on multiple timescales. Soc Neurosci Abstr 31:813.10.

Fusi S, Drew PJ, Abbott LF (2005b) Cascade models of synaptically stored memories. Neuron 45:599–611.

Gallistel CR (1994) Foraging for brain stimulation: toward a neurobiology of computation. Cognition 50:151–170.

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 Anim Behav Process 27:354–372.

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

Herrnstein RJ, Prelec D (1991) Melioration: a theory of distributed choice. J Econ Perspect 5:137–156.

Herrnstein RJ, Vaughan WJ (1980) Melioration and behavioral allocation. In: Limits to action: the allocation of individual behavior (Staddon JER, ed), pp 143–176. New York: Academic.

Herrnstein RJ, Rachlin H, Laibson DI (1997) The matching law: papers in psychology and economics. Cambridge, MA: Harvard UP.

Heyman GM (1979) A Markov model description of changeover probabilities on concurrent variable-interval schedules. J Exp Anal Behav 31:41–51.

Heyman GM, Luce D (1979) Operant matching is not a logical consequences of maximizing reinforcement rate. Learn Behav 7:133–140.

Hikosaka O, Nakamura K, Nakahara H (2006) Basal ganglia orient eyes to reward. J Neurophysiol 95:567–584.

Houston AI, Sumida BH (1987) Learning rules, matching and frequency dependence. J Theor Biol 126:289–308.

Huang Y-Y, Simpson E, Kellendonk C, Kandel ER (2004) Genetic evidence for the bidirectional modulation of synaptic plasticity in the prefrontal cortex by D1 receptors. Proc Natl Acad Sci USA 101:3236–3241.

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

Kawagoe R, Takikawa Y, Hikosaka O (1998) Expectation of reward modulates cognitive signals in the basal ganglia. Nat Neurosci 1:411–416.

Lau B, Glimcher PW (2005a) Caudate neurons encode both saccade and reward information in a free-choice task. Soc Neurosci Abstr 31:400.14.

Lau B, Glimcher PW (2005b) Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav 84, 555–579.

Lauwereyns J, Watanabe K, Coe B, Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. Nature 418:413–417.

Leon MI, Shadlen MN (1999) Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron 24:415–425.

Malenka RC, Nicoll RA (1999) Long-term potentiation—a decade of progress? Science 285:1870–1874.

Montague PR, Berns GS (2002) Neural economics and the biological substrates of valuation. Neuron 36:265–284.

O'Connor DH, Wittenberg GM, Wang SS-H (2005) Graded bidirectional synaptic plasticity is composed of switch-like unitary events. Proc Natl Acad Sci USA 102:9679–9684.

Otani S, Daniel H, Roisin M-P, Crepel F (2003) Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. Cereb Cortex 13:1251–1256.

Petersen CC, Malenka RC, Nicoll RA, Hopfield JJ (1998) All-or-none potentiation at CA3-CA1 synapses. Proc Natl Acad Sci USA 95:4732–4737.

Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. Nature 400:233–238.

Reynolds JN, Wickens JR (2002) Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw 15:507–521.

Reynolds JN, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. Nature 413:67–70.

Roitman JD, Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J Neurosci 22:9475–9489.

Rorie AE, Newsome WT (2004) The role of area LIP in a direction discrimination task with multiple reward contingencies. Soc Neurosci Abstr 30:20.11.

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

Schmolesky MT, Wang Y, Hanes DP, Thompson KG, Leutgeb S, Schall JD, Leventhal AG (1998) Signal timing across the macaque visual system. J Neurophysiol 79:3272–3278.

Schultz W (2000) Multiple reward signals in the brain. Nat Rev Neurosci 1:199–207.

Schultz W (2006) Behavioral theories and the neurophysiology of reward. Annu Rev Psychol 57:87–115.

Schweighofer N, Doya K (2003) Meta-learning in reinforcement learning. Neural Netw 16:5–9.

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

Shadlen MN, Newsome WT (1994) Noise, neural codes and cortical organization. Curr Opin Neurobiol 4:569–579.

Shadlen MN, Newsome WT (1996) Motion perception: seeing and deciding. Proc Natl Acad Sci USA 93:628–633.

Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J Neurophysiol 86:1916–1936.

Shidara M, Richmond BJ (2002) Anterior cingulate: single neuronal signals related to degree of reward expectancy. Science 296:1709–1711.

Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J Neurosci 13:334–350.

Soltani A, Wang X-J (2004) Exploring the neural basis of the matching law in choice behavior: a cortical network model with reward-gated learning. Soc Neurosci Abstr 30:668.14.

Staddon JER, Motheral S (1978) On matching and maximization in operant choice experiments. Psychol Rev 85:436–444.

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

Sugrue LP, Corrado GC, Newsome WT (2004b) Neural correlates of value in orbitofrontal cortex of the rhesus monkey. Soc Neurosci Abstr 30:671.8.

Sugrue LP, Corrado GS, Newsome WT (2005) Choosing the greater of two goods: neural currencies for valuation and decision making. Nat Rev Neurosci 6:363–375.

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.

van Vreeswijk C, Sompolinsky H (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. Science 274:1724–1726.

Vaughan W (1981) Melioration, matching, and maximization. J Exp Anal Behav 36:141–149.

Wang X-J (2002) Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36:955–968.

Williams BA (1988) Reinforcement, choice, and response strength. In: Steven's handbook of experimental psychology, Ed 2, Vol 2 (Atkison RC, Herrnstein RJ, Lindzey G, Luce RD, eds), pp 167–244. New York: Wiley.